SCUT sampling and classification algorithms to identify levels of child malnutrition

Juan Baraybar-Huambo[1] and Juan Gutiérrez-Cárdenas[2]

Universidad de Lima, Perú
jfbaraybar@outlook.com, jmgutier@ulima.edu.pe

Abstract. Child malnutrition results in millions of deaths every year. This condition is a potential problem in Peruvian society, especially in the rural parts of the country. The consequences of malnutrition range from physical limitations to declining mental performance and productivity for the individual. Government initiatives contribute to decreasing the causes of this disorder; however, these efforts are focused on long term solutions. The need for a fast and reliable way to detect these cases early on still exists. This paper compares classification techniques to determine which one is the most appropriate to classify cases of malnutrition. Neural networks and decision trees are used in combination with different sampling techniques, such as SCUT, SMOTE, random oversampling, random undersampling, and Tomek links. The models produced using oversampling techniques achieved high accuracies. Further, the models produced by the SCUT algorithm achieved high accuracies, preserved the behavior of the data and allowed for better representations of minority classes. The multilayer perceptron model that used the SCUT sampling techniques was chosen as the best model.

Keywords: Child malnutrition · Neural Networks · Decision trees · Random forest · Sampling techniques.

1 Introduction

Approximately 7.6 million children under five years old die every year. A third of these deaths are related to malnutrition [24]. In 2018, these deaths represented 45% of the total deaths and were mostly registered in low-income countries [25]. Some of the most common causes of this disease are nutritional deficiencies, low birth weights, the social and economic infrastructure, health, and sanitation services [17]. Malnutrition classification in children regards the detection and recognition of malnutrition, which could be directly related to diseases or illnesses with different causes. Malnutrition can result in deteriorated physical and mental conditions or even mortality. Malnutrition caused by illnesses can also be a recurrent factor in hospitalized children, but it could go undetected [16]. The consequences can be severe, including reduced physical and intellectual capacities, chronic diseases, infections, and higher mortality rates [11]. In 2017, malnutrition affected 12.9% of the children less than five years old [13]. As seen,

in the last couple of decades, this number has decreased significantly. Public policies implemented by governments have reduced this percentage; however, the results are seen in the long term [12].

This paper proposes the use of classification algorithms in combination with sampling techniques to predict malnutrition in children less than 5 years old in Lima, Peru. Since we faced unbalanced data, we decided to use sampling techniques such as SCUT [1], SMOTE [7] and Tomek links [23],before applying our classification techniques, which include the following: neural networks, decision trees and random forests. We noticed that by using sampling techniques, our classification accuracies are equal to or greater than those in the reviewed literature at the moment of this present research. For example, for the case of SCUT with a multilayer perceptron, we obtained an accuracy of 97.03% compared with the 77% in the literature (young to older women dataset).

2 Related Work

2.1 Factors associated with malnutrition

Mariños-Anticona,Chaña-Toledo, Medina-Osis, Vidal-Anzardo and Valdez-Huarcaya [14] determined that some of the most important factors at the national level were extreme poverty, low birth weights, and mothers' education levels. Additionally, lacking access to sanitation services and anemia were most relevant at the regional level. Sobrino, Gutiérrez, Cunha, Dávila, and Alarcón [21] analyzed the trends of malnutrition and anemia in children less than five years old from 2000 to 2011. The most dominant factors were the education level of the mother, living in rural areas, poverty, and having two or more kids. Chinchay [8] classified the different factors using macrocategories and determined that education, the lack of alimentary policies, the nutritional status of the mother, lactation period duration, and access to sanitation services were the most important. Bullón and Astete [5] determined that the most relevant factors for malnutrition in children less than three years old were the mother's education and the growth and development controls. Additionally, most of the malnourished children were from rural areas of the country, while the minority came from the capital.

2.2 Use of classification algorithms to predict malnutrition

Thangamani and Sudha [22] used supervised learning techniques to classify family health data and predict malnutrition. The data of 254 children were used, and the multilayer perceptron and random forest models provided the same accuracy (77.17%). Yu, Bhatnagar, Hogen, Mao, Farzindar and Dhanireddy [28] used a neural network to predict anemia in patients and determine the number of red blood cell packages that were needed for blood transfusions. Their neural network obtained an accuracy of 77%. Yılmaz and Bozkurt [27] developed an application to diagnose iron deficiency anemia in women using neural networks. They used 2,660 blood samples and obtained good results with the implemented

model (99.16% accuracy). Azarkhish, Raoufy, and Gharibzadeh [3] implemented a neural network and an adaptive neuro-fuzzy inference system (ANFIS) to predict iron deficiency anemia in women. They used a sample of 203 registries, and the best results were obtained by the neural network (96.29% accuracy), which was followed by the ANFIS (90.74%) and a logistic regression model (62.96%). Carcly et al. [6] used the CSA-kNN, CSA-Gini, FNN, and PNN algorithms to determine anemia. They used 2,600 blood samples from women, and the results were as follows. CSA-Gini achieved the highest accuracy (98.73%), and it was followed by the FFN (98.5%), the PNN (97.48%), and, finally, the CSA-kNN (96%). Aruna and Sudha [2] proposed a decision tree to predict malnutrition in women between 18 to 50 years old. The tree with the best performance was the logical decision tree (accuracy of 92.6%), which was followed by the multilayer perceptron (77.17%), random forest (77.17%) and ID3 (68.5%). Park, Kim, and Kyung [18] designed a model based on decision trees to detect the risk of malnutrition in the elderly. They used a dataset of 15,146 registries. The model with the best performance in the training phase was C5.0; however, CART had the best overall performance (78.1% training accuracy and 80.95% testing). Ye, Chen, Z., Chen, J., Liu, Zhang, Fan, and Wang [26] used decision tree analysis to determine the risk factors in a Chinese metropolis and compared the results to those of logistic regression models. They used data of 1091 children from six to twelve months old. They found that the classifier based on the decision tree was more accurate (88.8%) compared to logistic regression models (87.2%). Dalvi and Vernekar [9] used 500 images of red blood cells to compare the performance of different aggregation methods and classifiers in order to predict anemia. The stacking method that used a combination of the kNN and decision trees with naive Bayes was the best method (92.12% accuracy). Sanap, Nagori, and Kshirsagar [20] compared a J48 decision tree, a C4.5 one, and a support vector machine based on SMO. In the analysis, 514 blood samples were used. The results showed that the decision tree achieved an accuracy of 97.67% in the validation stage and was better than the support vector machine (87.35%). Markos, Doyore, Yifiru, and Haidar [15] applied data mining techniques to extract hidden patterns that allowed for the prediction of the nutritional status of children less than 5 years old in Ethiopia. The classifier based on the PART pruned rule induction obtained the best result (accuracy of 97.8%) and was followed by naive Bayes (97.6%) and J48 (97.3%).

3 Background

3.1 The SCUT Sampling Technique

The SCUT sampling technique was proposed by Agrawal, Viktor, and Paquet [1] to solve imbalanced datasets that are used for classification. There have been multiple studies on improving classification performance for binary class datasets, but little research has been conducted on multiclass balancing. This technique preserves the structure of the data without converting it, like other approaches tend to do (one versus one and one versus all). The technique uses

the SMOTE sampling technique combined with clustered undersampling to address between-class and within-class imbalances. Additionally, the expectation maximization (EM) algorithm was used. This algorithm replaces dense clusters using a Gaussian probability distribution. The SCUT algorithm proceeds as follows. The dataset is split into $\boldsymbol{n}$ parts, where $\boldsymbol{n}$ is the number of classes. The mean $\boldsymbol{m}$ of the number of instances of all classes is calculated. For all the classes that have several instances less than $\boldsymbol{m}$, oversampling is performed via SMOTE, thereby generating the necessary instances for the number of classes to be equal to $\boldsymbol{m}$. However, if the number of instances of the class is greater than $\boldsymbol{m}$, undersampling is performed using the EM technique to find the clusters and extract the same amount of instances of each. As a result, the behavior of these data is not lost, and the number of instances of the class is equal to $\boldsymbol{m}$. Finally, all the classes are merged in order to obtain a new dataset, and all classes have $\boldsymbol{m}$ instances. The algorithm proposed by Agrawal, Viktor and Paquet [1] is shown next:

```
Algorithm: SCUT
Input: Dataset D with n classes
Output: Dataset D' with all classes having m instances,
where m is the mean number of instances of all classes.

Undersampling:

For each D_i, i = 1, 2,. . ., n; where number of instances > m
    Cluster D_i using EM algorithm
        For each cluster C_i, i = 1, 2,. . ., k
                Randomly select instances from C_i
                Add selected instances to C_i'
        End For
        C = φ
        For i = 1, 2,. . ., k
                C = C ∪ C_i'
        End For
        D_i' = C
End For

Oversampling:

For each D_i, i = 1, 2,. . ., n; where number of instances < m
        Apply SMOTE on Di to get D_i'
End For
For each Di, i = 1, 2,. . ., n; where number of instances = m
    D_i' = D_i
End For
D' = φ
```

For i = 1, 2, . . ., k
D = $D^{'} \cup D_i^{'}$
End For
Return D_i

4 Methodology

We used the information from a survey that was conducted by the National Institute of Informatics and Statistics in Lima, Peru named The Demographic and Family Health Survey 2017 (http://iinei.inei.gob.pe/microdatos/index.htm). Approximately 20.355 registries were selected with information about children less than five years old. The variables that were used to build the models were the weight, height, sex, and anemia level of the children and the educational level of the mother. Based on the growth curves that were defined by WHO (www.who.int/child-growth/standards), the nutritional status of children was defined. Additionally, the Body-Mass-Index (BMI= of every patient was calculated. The implementation was done using Python, the scikit-learn, and imbalanced learning libraries. We must mention that the WEKA data mining tool was used to visualize information about the dataset. Feature encoding and normalization were applied when necessary, and multiple sampling techniques were used. The algorithms that were used were SCUT, random oversampling, SMOTE, random undersampling, and Tomek links. To classify the data, we generated models based on the multilayer perceptron, classification and regression tree (CART) and random forest models. The accuracy, sensitivity, specificity, and precision were used to measure the performance of the developed models. A cross-validation test was performed to ensure the reliability of the results of the given models. It is valuable to mention that when we loaded our model in WEKA, we found that the dataset was heavily unbalanced, presenting with three classes (Cronica=chronic, no tiene=absent, and aguda/moderada=acute/moderate) with different numbers of instances, as shown in Fig. 1 and Fig. 2.

To solve these problems, we applied data preprocessing techniques. First, data normalization was applied to continuous features such as weight and height. After that, categorical variables were encoded to be used for the future implementation of the models. Finally, to solve the unbalanced classes, the SCUT algorithm was applied. We found three main classes: chronic malnutrition (15,448 patients), no malnutrition or absent (4,185 patients), and moderate malnutrition (772 patients). The means were calculated for the 6,785 registries for each class. SMOTE was applied for the minority classes. In the case of the majority class, the EM algorithm was applied. When we used the elbow method, we found that the optimal amount of clusters inside the majority class was nine, as shown in Fig. 3. The number of instances that were randomly extracted from each cluster was 1,746. Once all classes were under or oversampled, they were united to form the new dataset that was going to be used for the implementation of the models.

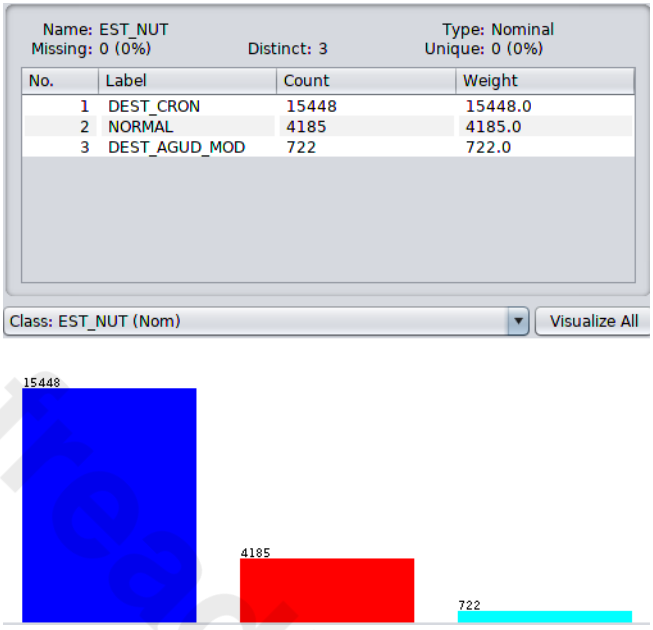

Fig. 1. Nutritional Status Imbalance.

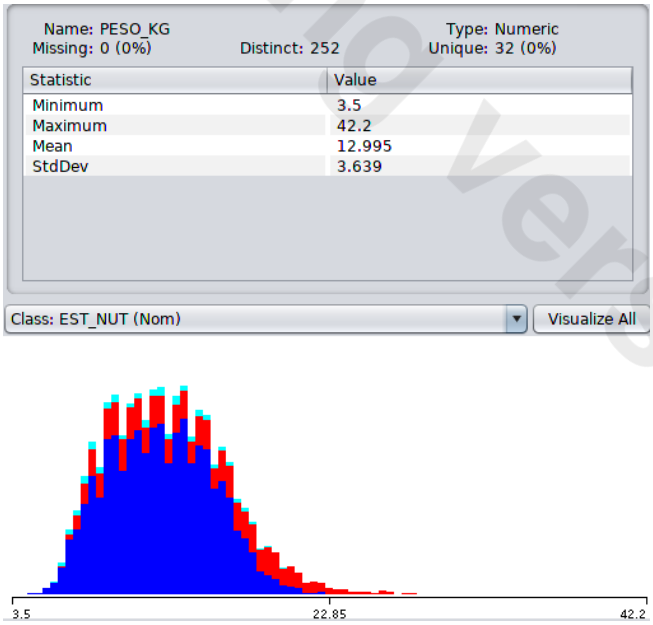

Fig. 2. Weight Distribution.

After the pre- processing stage, for example, the same weight data that we used before were transformed, as shown in Fig. 4.

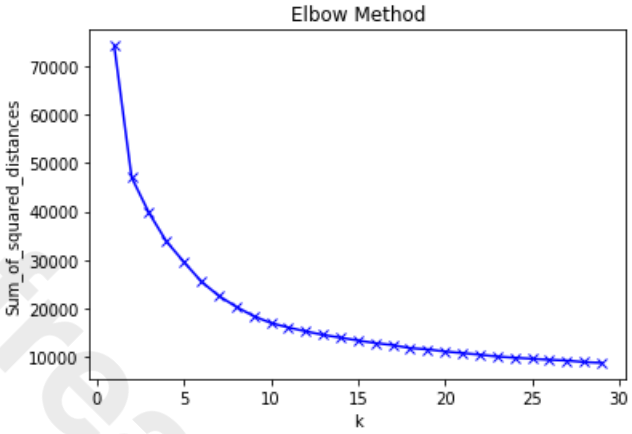

Fig. 3. Elbow method.

After the application of the SCUT algorithm, our balanced data classes are now as shown in Fig. 5. Additionally, after the application of SCUT, the dataset is entirely balanced and avoids under or overfitting problems. After that, we implemented classifiers based on multilayer perceptron (MLP), CART, and random forest. The hyper parameters for the MLP were a learning rate of 0.01, a hidden layer composed of 8 neurons, and a sigmoid activation function. The number of epochs was determined based on the decrease in the loss function. The optimal number of epochs was 250, as shown in Fig. 6. Iterations were performed to tune the hyperparameters from the different tree based-models for a set of various parameters. After that, we analyze the training and test set differences that were present and consider the accuracy of each model. Both the CART and random forest models used the entropy criterion for split selection. Additionally, the random forest model used 64 estimators.

5 Results

Based on the results shown in Table 1, the best combination is provided by the random oversampling technique and the Random forest or the CART algorithm. It is important to say that the undersampling techniques that were used did not converge for the MLP model with the given parameters. To validate our results, we used ten-fold cross-validation. The accuracies and standard deviations are shown in Table 2.

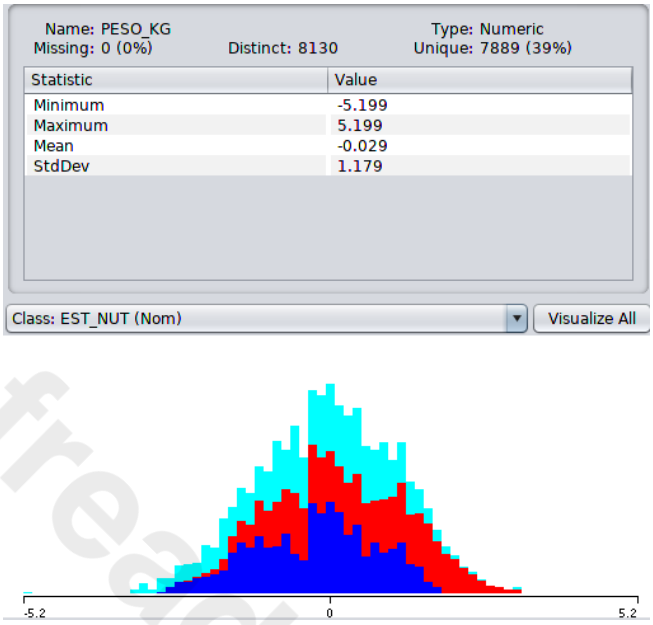

Fig. 4. Weight distribution after preprocessing.

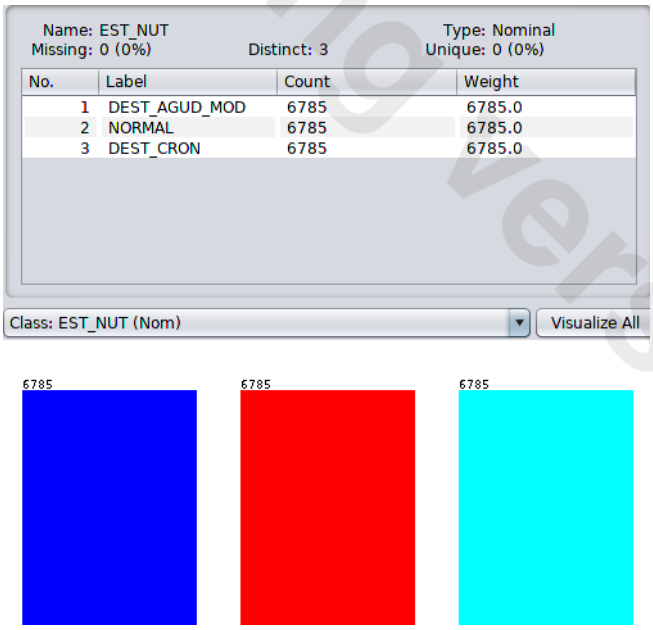

Fig. 5. Balanced classes after application.

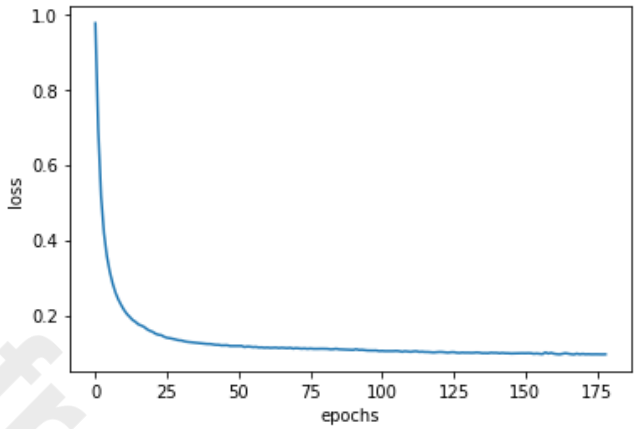

Fig. 6. Decrease in loss function by number of epochs.

Table 1. Performance of the classification algorithms using different sampling techniques in the testing phase.

SCUT				
	Accuracy	Sensitivity	Specificity	Precision
Multilayer Perceptron	97.03	97.03	98.52	97.02
CART	95.62	95.61	97.81	95.62
Random Forest	96.84	96.82	98.42	96.85
Random Oversampling				
Multilayer Perceptron	96.80	96.80	98.40	96.83
CART	99.05	99.05	99.53	99.06
Random Forest	99.05	99.05	99.53	99.06
SMOTE				
Multilayer Perceptron	97.69	97.69	98.84	97.71
CART	98.14	98.14	99.07	98.14
Random Forest	98.77	98.77	99.38	98.77
Random Subsampling				
Multilayer Perceptron	94.10	94.00	97.05	94.16
CART	84.69	84.61	92.36	84.60
Random Forest	85.98	85.85	92.96	86.44
Tomek Links				
Multilayer Perceptron	98.42	95.59	98.95	92.85
CART	96.99	89.52	97.56	90.59
Random Forest	96.36	82.01	95.73	96.44

Table 2. Ten-fold Cross-Validation results

SCUT		
	Accuracy	Standard Deviation
Multilayer Perceptron	96.12	1.71
CART	95.22	4.66
Random Forest	93.56	12.17
Random Oversampling		
Multilayer Perceptron	97.46	0.48
CART	99.39	0.22
Random Forest	99.20	0.29
SMOTE		
Multilayer Perceptron	97.54	0.73
CART	98.41	2.36
Random Forest	98.62	1.94
Random Subsampling		
Multilayer Perceptron	94.73	3.38
CART	86.42	4.01
Random Forest	88.27	3.68
Tomek Links		
Multilayer Perceptron	98.32	0.58
CART	97.17	0.73
Random Forest	96.28	1.08

6 Discussion

As we can observe in Table 2, the Random forest algorithm or the CART decision tree, in combination with random oversampling, provided surprisingly high results. However, we believe that these models may be overfitted because a large number of instances were randomly generated by the sampling techniques that were used. With respect to this overfitting, there is an agreement among researchers that the technique that is mentioned above is prone to this behavior due to the copying of the minority class samples [4]. The same deduction could be applied to the results that are generated by the SMOTE algorithm. This is because according to Agrawal, Viktor, and Paquet [1] , SMOTE, even though can avoid overfitting, it could generate this phenomenon in the presence of a highly imbalanced dataset [4].

However, the SCUT algorithm ensures the balance of the dataset and preserves the behavior of the original data. Methods such as SCUT are not prone to overfitting, so it is safe to use in the presence of imbalanced data [29, 4]. For the above reasons, when using the SCUT algorithm, the best chosen model to detect malnutrition was found to be the one that was generated by the multilayer perceptron, achieving an accuracy of 96.12% with a low standard deviation (+/- 1.71).

As a final test, we decided to compare our classification models by using the 5x2cv paired t-test proposed by Dietterich [10]. For this analysis, we select a

threshold of $\alpha = 0.05$, and a null hypothesis that states that both classifiers compared perform equally well. According to the results shown in Table 3, we can reject the null hypothesis in the comparisons of MLP and CART, and MLP and Random forest; meaning that choosing the MLP as a leading model would be suitable for this problem.

Table 3. 5x2cv paired t-test results

	MLP/Decision Tree	**MLP/Random Forest**	**Decision Tree/ Random Forest**
t-statistics	6.93373	5.31515	-1.94913
p-value	0.0009	0.0031	0.10879

Future work could be completed to test known combinations of techniques, such as SMOTE with Tomek links, that could perform well in the presence of skewed data [19] and to test the presented classifiers.

7 Conclusions

This research work explored the use of neural networks and decision trees combined with different sampling techniques to detect malnutrition in children from Lima, Peru. During the testing phase, the best result was from the random forest classifier and the fandom oversampling technique. A cross-validation test was performed, revealing that the SCUT algorithm as the best sampling technique overall, and the multilayer perceptron was one of the best classifiers. Future work could explore the use of other classification techniques, such as the naive Bayes and logistic regressions, to compare the overall performance of the presented classifiers and determine the best approach for detecting malnutrition in children less than five years old.

References

1. Agrawal, A., Viktor, H. L., & Paquet, E.: SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling. In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 1(lc3k), pp. 226-234. Lisbon, Portugal(2015). https://doi.org/10.5220/0005595502260234
2. Aruna, S., Sudha, P.: An Efficient Identification of Malnutrition with Unsupervised Classification Using Logical Decision Tree Algorithm. Research Journal of Pharmaceutical, Biological and Chemical Sciences, 4(2), pp. 365–373 (2016).
3. Azarkhish, I., Raoufy, M. R., Gharibzadeh, S.: Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. Journal of Medical Systems, 36(3), 2057-61.(2012). https://doi.org/10.1007/s10916-011-9668-3

4. Batista, G., Prati, R., Monard, C.: A study of the behavior of several methods for balancing machine learning training data ACM SIGKDD Explorations Newsletter, 6 (1) , pp. 20-29. (2004).
5. Bullón, C., & Astete, R.: Determinantes de la Desnutrición Crónica de los Menores de Tres Años en las Regiones del Perú: Sub-Análisis de la Encuesta Endes 2000. Anales Científicos, 77(2), 249. (2016). https://doi.org/10.21704/ac.v77i2.636
6. Çarkli Yavuz, B., Karagül Yildiz, T., Yurtay, N., Pamuk, Z.: Comparison Of K Nearest Neighbours And Regression Tree Classifiers Used With Clonal Selection Algorithm to Diagnose Haematological Diseases. AJIT-e: Online Academic Journal of Information Tech-nology, 5(16) (2014).
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, JAIR, 16 (2002), 321–357.
8. Chinchay, K.: Costos Económicos en Salud de la Prevalencia de Desnutrición Crónica en Niños Menores de 5 Años en el Perú en el Período 2007-2013. Lima (2015).
9. Dalvi, P. T., Vernekar, N.:Anemia detection using ensemble learning techniques and statistical models. In: 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1747-1751). IEEE (2016).
10. Dietterich TG.: Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, vol. 10(7), 1895–1923, (1998). https://doi.org/10.1162/089976698300017197
11. Garcia, L. S., Jave, C. M., Cárdenas, M. E. H., López, P. A., Sánchez, G. B.: Pobreza y desnutrición infantil. PRISMA ONGD (2002).
12. Instituto Nacional de Estadística e Informática: Desnutrición Crónica Infantil en niñas y niños menores de cinco años disminuyó en 3.1 puntos porcentuales, https://gestion.pe/economia/disminuye-desnutricion-cronica-infantil-pais-revela-inei-114809. Last accessed 6 July 2019.
13. Instituto Nacional de Estadística e Informática.: Desnutrición Crónica afectó al 12.2% de la población menor de cinco años de edad en el año 2018. https://www.inei.gob.pe/prensa/noticias/desnutricion-cronica-afecto-al-122-de-la-poblacion-menor-de-cinco-anos-de-edad-en-el-ano-2018-11370/. Last accessed 6 July 2019.
14. Mariños-Anticona, C., Chaña-Toledo, R., Medina-Osis, J., Vidal-Anzardo, M., Valdez-Huarcaya.: Determinantes sociales de la desnutrición cronica infantil en el Perú. Revista Peruana de Epidemiología, 18(1), 1–7. . (2014).
15. Markos, Z., Doyore, F., Yifiru, M., & Haidar, J.: Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey. J Health Med Inform, 5, 152. (2014).
16. Mehta NM., Corkins MR., Lyman B., B., Malone, A., Goday, PPS., Carney, LN., Monczka, JL., Plogsted, SW., Schwenk, WF.: American Society for Parenteral and Enteral Nutrition Board of Directors. Defining pediatric malnutrition: a paradigm shift toward etiology-related definitions. Journal of Parenteral and Enteral Nutrition, JPEN, 37(4),460-481. (2013).
17. Ministerio de Salud. Desnutrición Infantil Crónica y sus Determinantes de Riesgo. Lima, Lima, Perú (Marzo de 2010).
18. Park, M., Kim, H., Kyung, S.: Knowledge discovery in a community data set: Malnutrition among the elderly. Healthcare Informatics Research, 20(1), 30–38 (2014). https://doi.org/10.4258/hir.2014.20.1.30.

19. Prati R. C, Batista G. E, Monard M. C.: Data mining with imbalanced class distributions: Concepts and methods. Paper presented at the IICAI. (2009)
20. Sanap, S. A., Nagori, M., Kshirsagar, V. :Classification of anemia using data mining techniques. In International Conference on Swarm, Evolutionary, and Memetic Computing, pp. 113-121. Springer, Berlin, Heidelberg (2011).
21. Sobrino, M., Gutiérrez, C., Cunha, A. J., Dávila, M., Alarcón, J.: Desnutrición infantil en menores de cinco años: tendencias y factores determinantes. Revista Panamericana de Salud Pública, 35(2), 104-122 (2014).
22. Thangamani, D., Sudha, P.: Identification Of Malnutrition With Use Of Supervised Datamining Techniques –Decision Trees And Artificial Neural Networks. International Journal Of Engineering And Computer Science, 3(9), 8236-8241 (2014).
23. Tomek, I. Two Modifications of CNN. IEEE Transactions on Systems Man and Communications SMC-6(11), 769–772, (1976).
24. Wisbaum, W., Colaborado, H., Barbero, B., Allí, D., Arias, M., Benlloch, I., Lezama Isabel Tamarit, I. (2011). DESNUTRICIÓN INFANTIL: Causas, consecuencias y estrategias para su prevención y tratamiento. Unicef, 1, 21. (2011). url https://old.unicef.es/sites/www.unicef.es/files/Dossierdesnutricion.pdf. Last accessed 6 July 2019.
25. World Health Organization. Malnutrición urlhttps://www.who.int/es/news-room/fact-sheets/detail/malnutrition. Last accessed 6 July 2019.
26. Ye, F., Chen, Z., Chen, J., Liu, F., Zhang, Y., Fan, Q., Wang, L.: Chi-squared automatic interaction detection decision tree analysis of risk factors for infant anemia in Beijing, China. Chinese Medical Journal, 129(10), 1193-1199 (2016). https://doi.org/10.4103/0366-6999.181955
27. Yilmaz, Z., Bozkurt, M. R.: Determination of women iron deficiency anemia using neural networks. Journal of medical systems, 36(5), 2941-2945 (2012).
28. Yu, C. H., Bhatnagar, M., Hogen, R., Mao, D., Farzindar, A., Dhanireddy, K.: Anemic Status Prediction using Multilayer Perceptron Neural Network Model. In: 3rd Global Conference on Artificial Intelligence, GCAI, pp. 213-220. (2017).
29. Zhuoyuan Zheng, Yunpeng Cai, Ye Li: Oversampling method for imbalanced classification, Computing and Informatics, 34(5), pp. 1017-1037. (2015)